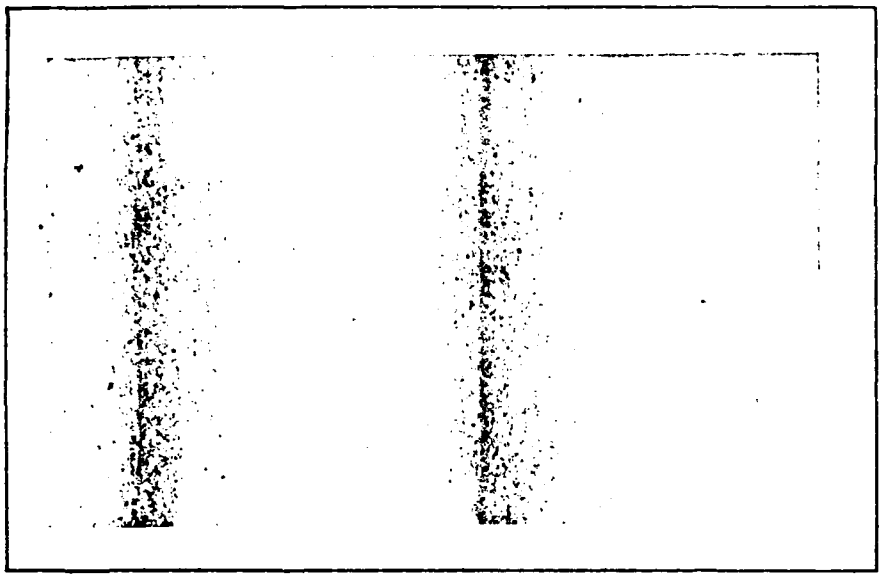
AD A121925



DEPARTMENT OF STATISTICS

DTIC FILE COPY

DTIC ELECTE NOV 30 1982 S D E

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

This ... been approved for public release and sale; its

82 11 29 072

Best
Available
Copy

8

SPECIFICATION AND IMPLEMENTATION OF AGE, PERIOD AND COHORT MODELS

by

Stephen E. Fienberg*

and

William M. Mason**

*Departments of Statistics and Social Science
Carnegie-Mellon University
Pittsburgh, PA 15213

**Population Studies Center and Department of Sociology
University of Michigan
Ann Arbor, MI 48109

Technical Report No. 235

January, 1982

DTIC ELECTE NOV 30 1982 E

The preparation of this paper was partially supported by the Office of Naval Research Contract N00014-80-C-0637 at Carnegie-Mellon University. Reproduction in whole or part is permitted for any purpose of the United States Government.

This document has been approved for public release and sale; its distribution is unlimited.

SPECIFICATION AND IMPLEMENTATION OF AGE, PERIOD AND COHORT MODELS*

by

Stephen E. Fienberg
Departments of Statistics and Social Science
Carnegie-Mellon University

and

William M. Mason
Population Studies Center and Department of Sociology
University of Michigan

January 1982

*This paper is a revision of material presented at the Conference on Analyzing Longitudinal Data for Age, Period and Cohort Effects, sponsored by the Committee on the Methodology of Longitudinal Research of the Social Science Research Council, held at Snowmass, Colorado, June, 1979. We are indebted to Michael Meyer for helpful discussions on degrees of freedom and interactions. Carol Crawford ably "wordprocessed" the manuscript using Textedit, while at the same time learning that system. We trust that she did not age too much, nor wishes she were a member of another cohort in a different era. Preparation of this paper was supported in part by Grant SES 80-08573 from the National Science Foundation to the University of Minnesota; Office of Naval Research Contract N0014-80-C-0637 at Carnegie-Mellon University; National Science Foundation grants SOC 78-17407 and SES 81-12192 to the University of Michigan; National Institute of Child Health and Human Development Grant 1R01 HD15730-01 to the University of Michigan. Reproduction in whole or part is permitted for any purpose of the United States Government.

I. INTRODUCTION

For the past 80 years or more, social scientists have attempted to analyze cross-time data, using as explanatory variables age and time (or phenomena that are time-specific). When such data are analyzed in aggregate forms, age and time are typically grouped and polytomized. More recently, some investigators have adopted an analytic focus in which cohort membership, as defined by the period and age at which an individual observation can first enter an age-by-period data array, is held to be more important than age or period for substantive understanding. This focus has led to age-cohort and period-cohort models, as distinguished from age-period models.

This paper is concerned with models for situations in which all three of age, period, and cohort are potentially relevant for the study of a substantive phenomenon. Relevance can be manifest for a number of reasons. First, the process under examination may be thought to depend on all three of age, period and cohort; in this case theory requires formulation of a model in which outcomes are determined jointly by age, period and cohort. The paper in this volume by Mason and Smith on tuberculosis mortality provides an example of the relevance of the full age-period-cohort specification. Second, the process studied may be thought by some investigators to depend on any one or two of age, period and cohort but not all three, and by other investigators to depend on a different proper subset. Here

the age-period-cohort (APC) specification may also be appropriate so long as competing reduced models (in the sense of Fienberg and Mason, 1978) are compatible.[1] Third, a single investigator may have in mind a specific reduced specification, but would want to fit the full APC model for baseline comparisons. The education example in Fienberg and Mason (1978) illustrates this case. Fourth, the APC model can serve as the starting point for more complicated modelling efforts involving the addition of specific interaction terms. We explore this latter approach in some detail.

There is, in a sense, a possible double parsimony associated with APC models. They can involve fewer parameters than say an age and cohort plus interactions model. There can also be an epistemological parsimony associated with such models since they require alternative conceptualization rather than extensions to handle ad hoc interactions, as might be the case with two-variable perspectives. Adding parameters to capture pieces of interaction between, say, age and period not reflected in the additive interaction terms (i.e., "main effects") for cohort is not precluded. Mason and Smith (this volume) illustrate such an approach in dealing with anomalous results in the fit of APC models to tuberculosis mortality rates.

DTIC COPY INSPECTED

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A

We further narrow our focus in this paper by an emphasis on accounting specifications, using as our accounting categories age, period, and cohort. As with their "green-eyeshade" counterparts, these models do not explain so much as they provide categories with which to seek explanation. For accounting models to have value, the parameterizations of the general framework must be linked to phenomena presumed to underly the accounting categories. This is a conceptual linkage minimally, and maximally an empirical linkage as well. If measurement of the underlying phenomena is possible, the accounting categories can be dispensed with. But if the accounting categories are superfluous, then the variety of problems we touch on in this paper are generally, though not always, of considerably less interest. For example, the identification problem created by the linear dependency of age, period and cohort is irrelevant if underlying empirical measures are available for any one of age, period or cohort (e.g., a business cycle measure for period, or cohort size for cohort membership). On the other hand, the design issues we touch on are pertinent regardless of the wealth of information that may be available for measurement purposes.

Accounting models have long been objects of attention, largely because of the conundrum posed by the identification problem. This problem is solved and relatively well understood by now (Fienberg and Mason,

1978). In our view it is appropriate to shift attention to other problems that arise in cohort analysis, and this we do in the present paper.

Using our earlier paper as a point of departure, this essay focusses on a number of points that have been considered problematic. We begin in Section II with a discussion that relates modelling efforts to historical and universal processes. Of interest here are arguments over the relevance of statistical inference. We also discuss the relevance of the group or macro level perspective afforded by an APC model, and provide a framework in which to understand the complications introduced by data aggregation. Section III reviews the various data structures that have been accompanied by cohort based models, and Section IV reviews the technical details of the APC accounting framework. To the APC description of Fienberg and Mason (1978) we add a continuous-time polynomial representation, and we conclude that the identification problem in APC models is inescapable--no matter which representation we choose to work with.

Fienberg and Mason (1978) adopt the rather dogmatic view that age-period, age-cohort, and cohort-period effects are virtually inestimable when added to the basic APC model. This is true only in a narrow sense, and Section V presents extensions of the APC model which allow for restricted

interactions. In most of the extensions the estimability of the interaction terms hinges primarily on dimensionality, and not identification and specification arguments.

II. ON DEFINING THE PROBLEM

A. Scope

The planning of cohort studies, and discussions of various approaches and study designs, can be aided by a framework for thinking about (a) the purposes of the analysis, (b) the nature of the data batch, and (c) the relevance and role of the distinction between micro and macro levels of analysis. The remarks we offer here are not limited in their relevance to APC analysis, and are not new, although their elaboration in this context may be unusual. They are, moreover, equally applicable to APC models in their simplest form, more complex models, reduced models, and models in which some or all of the accounting dimensions are replaced by more direct measures of the substantive phenomena presumed to underlie the dimensions.

To begin with, we distinguish between different potential analytic goals. The polar extremes often invoked are (i) the use of cohort analysis to organize historical material surrounding a particular era, place and set of events, and (ii) the use of cohort analysis to abstract from historical context to some more general universe. An example of the former might be Ryder's (1965:849) mention of the rise of the Bourbaki group in French academic mathematics as a consequence of the mortality induced by World War I. The example is highly time and place specific, and cohorts are invoked to clarify the composition of the innovating group known as Bourbaki. No particular model is

supposed by this citation, but one could envision coding mathematicians by date of birth (or date of degree), age (or time since degree), and choice of subject matter. The resultant data structure could be analyzed from the standpoint of an APC specification or some other, more complex, formulation. The researcher could use the results to describe accomplishments in a particular era, but it might also be justifiable to interpret them as indicating something more universal--perhaps the function of wars as instrumentalities for increasing or inducing creativity. The researcher's intent as to generalizability is rarely clear from specific analyses.

Lack of clarity regarding the broader goals of the analysis engenders confusion about the usefulness of the results. On one hand, to those most concerned with historical specificity, the abstraction imposed by a statistical model looks like neglect of substantive information about the setting in which the data are embedded. On the other hand, to those most interested in generalizability, some sacrifice of historical detail is essential: It is inherently difficult to accommodate historical richness in statistical models, and it is undesirable to attempt to do so to any considerable degree. To complicate matters, there remains the further problem of deciding whether the model employed is reasonable, given the supposition that modelling of the substantive phenomenon is in principle meaningful. Determining whether a model is

reasonable is not always separable from the question of abstraction or focus. For example, whether to include interactions above and beyond the "additive" components of an APC specification might hinge on the analyst's sense that the goal is to generalize rather than to fit certain obtrusive facets of the data.

In thinking about the focus of the modelling effort, it is helpful to consider the nature of the data batch. Are the data fruitfully thought of as essentially all that are conceivably available for analysis (as might be the case, for example, in working with census data), or can the data be thought of as a sample? Thinking about the nature of the data batch in conjunction with the focus of analysis suggests the following classification:

	Data Batch	
	All	Sample
Historical, Descriptive Emphasis...	(1)	(2)
Universal, Processual Emphasis.....	(3)	(4)

Cell (3) is empty: By definition, if your interest is in generalizing from the data you can not have it all. Even if you were working from perfect census data, your batch would be a sample in time. The remaining cells are nonempty, and the problem of determining how to treat a given study must be addressed in these instances. Whether a study belongs in

cell (1), (2), or (4) seems to depend on how one wishes to think about the discrepancies between observed data and fitted values.

In general, we are concerned with modelling efforts in which expressions such as

$$Y = \text{fit} + \text{residual}$$

are employed. Expressions of this kind depend on some prior conceptualization of the phenomenon under study. Whether they are informative depends on abstract evaluation of the conceptualization, but also on realized values of the fitted terms and residuals. The residuals, or discrepancies between the observed response variable and the numerical values fostered by the modelling process, can have any of the following sources:

(i) Sampling accidents (discrepancy between sample and finite population or super-population).

(ii) Stochastic fluctuation for individuals or groups (this can be present even if there are no sampling errors). The system is subject to shocks which may or may not be treated as "random."

(iii) Curve fitting errors: If the model is wrong the "predictions" are likely to be wrong.

Now, what is the linkage between cases (1), (2), (4) and the classification of ways to think about residuals? First, if you think of residuals as arising solely from specification errors, then you can only be concerned with historical description, and you have to think that you have all of the data relevant to the problem. Second, if you

think of the residuals as due to sampling accidents, then you clearly can not think of the data batch as consisting of all the data, but you could have either a historical or a more abstractive interest. That is, thinking of your data as a sample does not force a decision as to emphasis, but it does require consideration regarding whether inference is to a finite population or to a super-population (Hartley and Sielken, 1975). Third, if you think of the residuals as reflecting stochastic fluctuation but not sampling accidents, then you are again thinking in terms of historical, noninferential analyses. You are saying you have all the data and the right model. Your model does not capture every nuance of historical reality, but you do not intend it to--that would contradict the parsimony sought by modelling.

Clearly there is no logical ordering here. You could start by saying that you were interested in historical description and that you thought you had all the data. From this you would be drawn to thinking of residuals as indicating either stochastic fluctuation of no interest to you, or indicating curve fitting errors.

Given that one conceives of the data batch as some kind of sample, it is by no means obvious that a plausible super-population will come to mind, and it is certainly not the case that standard errors will be readily computable. Thus, taking the stance that inference is desirable is far

removed from being able to carry out the inference. When it is not possible to carry out the inference, what can be the role of statistical modelling?

When the basis of inference is unclear it seems helpful to think of the results of statistical modelling as providing windows. What we gain is a view of the world from a particular location. Shifting our view by applying another model amounts to looking at the data through a different window. We have experience in the interpretation of statistical models when there is a reasonably good mesh between data and assumptions. The panoply of numerical information from statistical models tells us what the world would look like if the assumptions were met. We may decide that the view is a helpful one, or that it is so implausible that the "as if" perspective can not be maintained: The data depart too much from the underlying assumptions of the statistical model employed.

B. Levels of Analysis

Historically, the conceptual attractiveness of cohorts as analytic differentia has been that they refer to sets of individuals with shared experiences. The question is sometimes raised whether APC models describe individual behavior, group behavior or a combination of the two. This question is rooted not so much in a concern for when a set of individuals can be said to form a group, but rather owes its currency to a tendency on the part of some discussants to link the level of aggregation in a data set to a

conceptual unit of analysis. There is, for example, a suggestive power to "grouped data," as in averages or other summary measures, conditional on age and period.

If the unit of analysis is at issue, then the nature of explanation compatible with APC models is no less so. In this subsection we advance the view that cohort analysis of any kind is inherently multi-level. The statement of this position forms one answer to the question of the appropriate unit of analysis. A second answer is attendant upon the multi-level conception: Given a perspective on the desired conceptual units of analysis, it becomes possible to assess the impact of data aggregation. In the next subsection we show that common forms of data aggregation can lead to biased estimates of effects in APC models, or expanded models.

To the extent that there is anything distinctive about APC models or cohort analysis more generally, with respect to the conceptual units of analysis, it is that the notion of cohorts when used in the social sciences frequently carries with it the tacit assumption that the object of study is not context free. Given this assumption, adequate APC models can not simply reflect formulations of behavior at a single level or aggregation or analysis, and must instead be sustained by a broader multi-level perspective.

Within a multi-level perspective (Mason, 1980) group or macro level phenomena can have an impact on individual or micro level phenomena.[2] To test theoretical understanding of group phenomena it is necessary to specify how the macro variables are translated into micro impacts. Doing so requires a theoretical conception of how specific macro variables might affect certain micro variables, and it requires also that the linkages between levels actually be measured. Thus, for a macro phenomenon to have an impact on the endogenous variable, its force must be realizable through variables measured at the micro level. This is an ultimate test of a macro theory. When we estimate APC models, we are unlikely to be able to develop such tests, given data restrictions. APC specifications are incomplete because of this, but then so are the alternatives. The acid test described here is a goal rarely attained, but its pursuit will promote clarity in conceptualization and explanation.

The goal of articulating linkages across levels of organization is not reductionist, and it is not unusual. It is precisely this linkage which remains to be determined before the scientific case against cigarette smoking can be closed (Brown, 1978; U.S. Department of Health, Education and Welfare, 1979). In quantitative statistical analyses attention to a total logical structure including cross-level linkages requires the specification of variables which measure how relevant macro phenomena are translated into

micro experiences. Cohort size, for example, is one of the most commonly cited "mechanisms" by which cohorts are differentiated. This variable has been hypothesized to be responsible for variation in a remarkable array of phenomena, including employment rates, wage rates, fertility, divorce, crime, and political alienation (Easterlin, 1978). A full statement of any cohort model which attributes the impact of cohort membership to cohort size must, in our view, include specification of the mechanisms which lead to the hypothesized consequences and designate variables presumed to be indicators of the mechanisms involved.

The multi-level perspective we advocate is generally applicable across the entire spectrum of cohort models, not just to APC specifications. To summarize this discussion, we are proposing the following: (a) When used, the accounting categories (age, period and cohort) must be well justified. This means that the analyst must be able to conceptualize how the accounting categories might be replaced with substantive variables, and that upon doing so the resulting specification must retain plausibility. (b) The interpretations intended by the use of the substantive variables must be, in principle, verifiable. That is, it should be possible to state the intervening steps between the macro level and the micro level, even if

data are unavailable to verify the interpretation in full detail. (c) When possible, the verification should be attempted.

C. Data Aggregation

A parallel is sometimes drawn between the degree of aggregation in a data set and the conceptual unit of analysis, so that, for example, the use of aggregated data is held to necessitate a conceptual analytic focus with similarly aggregated units of analysis. There is no such ready connection between the conceptual units of analysis and the degree of aggregation in the data. Nonetheless, once conceptual units of analysis have been selected, the relevance of data aggregation can be studied. This is no less true for models in other contexts, but we shall consider the implications of the form of the data for APC and related models since questions about aggregation seem especially persistent for them.

The potential deleterious effects of aggregation that most concern us are those of coefficient bias. We consider the impact of data aggregation on coefficient bias with the aid of a general equation which links a response variable to age, period, and cohort in discrete form, and to micro level explanatory (M_r) variables, macro level global explanatory (G_s) variables, and micro level aggregate explanatory variables ($\bar{M}_r$) constructed from micro variables. A global variable is not decomposable into a corresponding micro level variable, whereas an aggregate variable is (Lazarsfeld

and Menzel, 1961). The equation is considerably more general than would be needed were we to restrict our attention solely to APC specifications. This is appropriate, since in many applications it is possible to move beyond the simplest case of APC modelling to include other variables or to substitute for one of the accounting dimensions. In the equation, g is an operator over the row space of Y, and f is an operator over the column space of the predictors, that is, it defines the functional form relating g(Y) to the predictors. Consider

$$(1) \qquad g(Y_{ijk\ell}) = f(A_i, P_j, C_k, M_{ijk\ell r}, G_s, \bar{M}_r)$$

where

$i = 1,\ldots,I$ subscripts age

$j = 1,\ldots,J$ subscripts period

$k = i + j - 1 = 1,\ldots,K$ subscripts cohort

$\ell = 1,\ldots,n_{ij}$ subscripts micro observations within age-period combinations

$r = 1,\ldots,R$ subscripts (micro) M-variables, that is, variables for which values can vary over ℓ within age-period combinations

$s = 1,\ldots,S$ subscripts (global) G-variables (which cannot vary over ℓ within age-period combinations)

$\bar{M}_r$ = the aggregation of the rth micro variable (so that M_r does not vary over ℓ within age-period combinations)

We will use equation (1) to catalogue a variety of possibilities with respect to data aggregation. The basic question I consider is: Under what conditions, defined by

combinations of assumptions about g and f, will coefficient estimates be biased, relative to a baseline or fundamental specification? This fundamental specification is one in which g is the identity transformation. That is, we suppose the comparisons are best made with the situation in which the analyst has all of the possible relevant information.

To generate the catalog we allow g to be either linear or nonlinear, and cross these possibilities with the parallel dichotomy for f. We then use the resulting four combinations to consider bias in the coefficients of the predictors of equation (1).

If g is a linear operator, we might think of the following standard situations in which the analyst has available

$$\bar{Y}_{ijk\bullet}, \quad n_{ij}, \quad \text{and/or} \quad s^2_{Y_\ell|ij}$$

for the ij combinations. This covers the standard case of a dichotomous response variable, since in that instance $\bar{Y}_{ijk\bullet}$ is a proportion, and the conditional binary counts are recoverable from the information given. The generalization to a polytomous response variable is straightforward. If g is a nonlinear operator, we might think of the ij conditional medians. Saying that f is linear is to indicate that f is linear in the parameters, but not necessarily in the predictor variables. If f is nonlinear, then f is nonlinear in the parameters. An operator h may, or may not exist which would linearize the relationship. Table 1 summarizes the outcome of our consideration of aggregation

bias relative to the case in which all of the micro data are available. It has been arrived at by application of elementary ideas concerning aggregation and the analysis of covariance.

The results summarized in Table 1 suggest that even under the simplest form of aggregation, in which both **g** and **f** are linear operators, some, but not all, effects have or can have unbiased estimators. We will first review the results for **g** linear, and then take up the case in which **g** is nonlinear. To begin with, suppose we consider just the age, period and cohort coefficients. If the only available response variable is represented as a set of ij conditional means, and if the n_{ij} or estimated conditional variances are available, <u>and</u> if **f** is linear, then it is possible to obtain unbiased estimates of the age, period and cohort effects (case i). On the other hand, suppose **g** defines conditional averages, and **f** defines an exponential function (case ii). Then the effects will in general be biased, since the mean of a set of logarithms is not equal to (or a simple function of) the logarithm of a mean. The exception to this result occurs when the response variable is inherently discrete. In that case we presume sufficient information exists to resolve the conditional means (proportions), say, back into the dichotomous or polytomous counts. In this instance it is possible to obtain unbiased estimates for the age, period and cohort effects.

[illegible] the effects of data aggregation on coefficient bias, for APC and more general models

	f is linear	**f** is nonlinear
	g is linear	
[illegible]	i) unbiased	ii) biased unless Y is discrete with recoverable ij counts
[illegible]	iii) biased unless **g** operates within distinct combinations of values of M-variables	iv) biased unless **g** operates within distinct combinations of values of M-variables and Y is discrete with recoverable ij counts
[illegible]	v) unbiased if the G-variables vary between ij categories but do not within	vi) biased unless Y is discrete with recoverable ij counts
[illegible]	vii) biased unless $\bar{M}$ is a valid proxy of a G-variable	viii) biased unless $\bar{M}$ is a valid proxy of a G-variable and Y is discrete with recoverable ij counts
	g is nonlinear	
[illegible]	ix) biased	x) see ii
[illegible]	xi) biased	xii) see iv
[illegible]	xiii) biased	xiv) see vi
[illegible]	xv) biased	xvi) see viii

Note: This table [illegible] if Y is discrete, then the functional form of choice (**f**) is nonlinear.

To say that it is possible to obtain unbiased estimates does not imply that they are automatic. In case i, application of ordinary least squares when the response variable is treated as a set of cell means will provide biased coefficient estimates. Weighted least squares, using the n_{ij} or $s^2_{Y_\ell|ij}$, will yield unbiased estimates. Likewise, if the response variable is inherently discrete, it is not enough to know this and then to regress conditional proportions on age, period and cohort. This will lead to biased coefficients for more than one reason. Rather, one would usually want to recover the dichotomous or polytomous counts within ij-combinations, so that the full information available from the data could be used (case ii).

Next consider the coefficients of the micro (M) variables, assuming to begin with that f is linear. If the operation of g is such that the $Y_{ijk\ell}$ cannot be related to the $M_{ijk\ell r}$, then the estimator for the effects of the M-variables will be biased (case iii). Suppose that we have available the $\bar{Y}_{ijk+}$ and the underlying response variable is not discrete. Then the estimable relationship between the $\bar{Y}_{ijk+}$ and $M_{ijk\ell r}$ amounts to a relationship between means (i.e., between $\bar{Y}_{ijk+}$ and $\bar{M}_{ijk+r}$), and the regression line through the means is not necessarily the same as the regression line through the micro data. An exception occurs if the operation of g is such that the $Y_{ijk\ell}$ can be related to the $M_{ijk\ell r}$. This will happen when the operation of g is conditional on the distinct combinations of values of the

M-variables. Suppose, for example, that race and sex are the M-variables. Then if the response variable consists of $\bar{Y}_{ijk+}$ conditional on the race-sex combinations, all micro information is preserved. If f is nonlinear, the only condition under which the M effects are unbiased occurs when Y is discrete with recoverable dichotomous or polytomous counts, and the underlying recoverable tabulation is $Y \times A \times P \times M_1 \ldots \times M_r \ldots \times M_R$ (case iv).

Estimating the G coefficients is relatively unproblematic, if g and f are linear operators. Under these conditions (case v), the effects will be unbiased if the level of aggregation is such that it is natural for values of the G-variables to vary across ij categories, but not within them. Given this degree of aggregation, the space spanned by the G-variables is a subset of the space spanned by the age, period and cohort accounting categories. If f is nonlinear (case vi) this result is no longer correct, unless the response variable is discrete and the ij-conditional dichotomous or polytomous counts are recoverable. This is, of course, a common situation, as when the data consist of percentages and base n's, conditional on age and period, and it also makes sense to define global variables as varying over age-period categories.

The coefficients of the $\bar{M}$-variables (variables which are linear aggregations of micro variables) will be unbiased if both g and f are linear, provided that the $\bar{M}$-variables

are valid proxies of unmeasured G-variables. If the $\bar{M}$-variables are included only because the $M_{ijk\ell r}$ are unavailable, however, the coefficients of the $\bar{M}$-variables will be biased (case vii). The same conclusion holds if f is nonlinear (case viii) with the additional stipulation that Y must be discrete, with recoverable ij-conditional dichotomous or polytomous counts. Thus, the conclusions we reach concerning $\bar{M}$-variables are identical to those for G-variables provided that the $\bar{M}$-variables are valid substitutes for the G-variables. The $\bar{M}$-variables would generally not be so regarded if they were included only because the $M_{ijk\ell r}$ were unavailable. If this were the reason for inclusion, then the conclusions of cases iii and iv would apply.

Finally, we consider the possibility that g is nonlinear. When this condition holds, then all coefficients will generally be biased. Exceptions to this conclusion occur when Y is discrete with recoverable counts as specified for cases x, xii, xiv and xvi. These exactly parallel cases ii, iv, vi and viii, respectively. As an example, suppose that the analyst is given ij conditional logits and the n_{ij}. Then it is possible to recover the ij-conditional binary counts, and thus possible to obtain unbiased estimates of the age, period and cohort effects in a logit model.

The preceding exercise shows that there are forms of data aggregation which entail biased estimation of effects. There are also forms of data aggregation which permit unbiased estimation, and, depending on the nature of the response variable, amount to nothing more than compact representation of the data with no loss of information. It is a simple error, commonly made, to assume that because the data are in tabular form the data have been aggregated. This need not be the case. Adopting a multilevel perspective, as we have done here, permits the recognition that having full information rests on knowing the $Y_{ijk\ell}$ and the $M_{ijk\ell r}$ (if relevant). Given a decision as to the meaning of full information, it is possible to assess the impact of data aggregation on coefficient bias. The degree of aggregation poses a complication which is separate from the choice of units of analysis.

III. DATA STRUCTURES

When social and demographic researchers write about cohorts and cohort analysis, they typically have in mind a particular form of data array, and a design for collecting the relevant information. In this section, we discuss the six major data structures used in cohort analyses, and related aspects of design and data collection. We begin with the simplest structure and build towards the more complex, noting as we pass which structures link naturally to cross-sectional data collection, and which to longitudinal data collection. Some structures are compatible with both longitudinal data collection and repeated cross-sections, and we so note. To make the pictures of the data structures comparable across types we adopt the convention of letting periods or time vary horizontally, and cohorts and age vary vertically.

A. Single Cross-Section Studies

If we collect data by means of a single cross-sectional survey, then we can structure them by age groupings as in Figure 1 to represent a single synthetic cohort.

This structure is most useful when it can be assumed that cohort and period effects are known, or known to be nonexistent. When this is true, cohort differences in a single cross-section can be treated as age differences, or at least can be manipulated to reflect age differences.

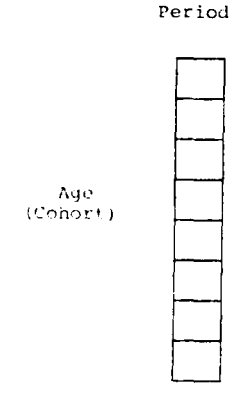


Figure 1. Synthetic Cohort From A Single Cross-Section

Using this approach for two or more cross-sections substantially separated in time leads to multiple synthetic cohorts. These have been used in demographic settings.

Of course, we could approach a single cross-section somewhat differently, assuming that age and period effects were known, or known to be nonexistent. Then the age differences could be treated or manipulated to reflect cohort differences.

One problem with the single cross-section study is that the kind of prior knowledge indicated above is usually not available. Hence applications are based on approximations to this particular ideal. A second problem, and a serious one, is that engendered by selectivity. Suppose that we are interested in cohort fertility and choose to examine only the fertility of women no longer in the childbearing age range--and that our data come from a single cross-section. Then we immediately run into the problem of differential mortality by cohort, and this creates selectively different samples by cohort. A different example encounters selectivity at the opposite end of the age distribution. Mare (1979) examines probabilities of continuing to a next level of education on a cohort by cohort basis, using the Occupational Changes in a Generation II data (Featherman and Hauser, 1975). For data of this kind, the selectivity encountered is that the younger the

cohort, the less likely individuals are to have completed their education. Hence, comparisons between cohorts that are insensitive to this problem are likely to be biased.

B. Studies of a Single Cohort Followed Over Time

As the name of this data structure suggests, it is longitudinal in form. In this instance there are multiple time points for a single cohort, as opposed to multiple cohorts observed at a single point in time.

This design is especially useful for analyzing a developmental process, and requires assumptions parallel to those for a synthetic cohort--in this case that period and cohort effects are known or known to be nonexistent. Clearly, the analyst might wish to be able to follow more than one cohort, but matters of cost are always relevant. One example of such a study is that of Wolfgang, Figlio and Sellin (1972), which follows the criminal careers of all males born in 1945, residing in Philadelphia from age ten to age eighteen. In fact, very recently Wolfgang and his collaborators have followed up the original study with a second cohort of males born in 1958 (Wolfgang, 1982).

C. Multiple Cross-Section Studies

This data structure is based on independent replicated cross-sectional surveys, or an aggregate cross-sectional population data. The resulting data are typically displayed in the form of a rectangular age by period table

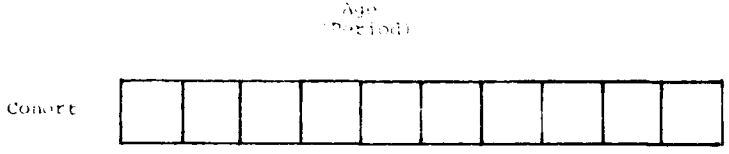


Figure 2. A Single Cohort

as in Figure 3, but as Fienberg and Mason (1978) note, they can also be displayed in a parallelogram-shaped age by cohort or period by cohort table.

If age groups and period have the same spacing, then subsamples in the same cohort can be linked across surveys (periods) as indicated by the dotted lines in Figure 3. This is consequently the first design discussed thus far which admits the possibility of calculating age effects adjusted for period and cohort differences; and cohort effects adjusted for period and age differences. Indeed this is the first design we have considered which does not require us to assume that some of the potential effects are known to be nil.

There are abundant examples of this form of data, analyzed from a cohort perspective (e.g., Greenberg, Wright, and Sheps, 1950; Carr-Hill, Hope and Stern, 1972; Farkas, 1977; Pullum, 1977, 1980; Fienberg and Mason, 1978; and Mason and Smith (this volume)).

The design is not without difficulties. First, there is an identification problem, assuming that we use an APC accounting framework. This problem arises from the dependency between age, period and cohort. Its solution depends on prior knowledge, reasoning about the process under scrutiny, or theory more generally. In many examples, there is no coherent body of substantive reasoning which

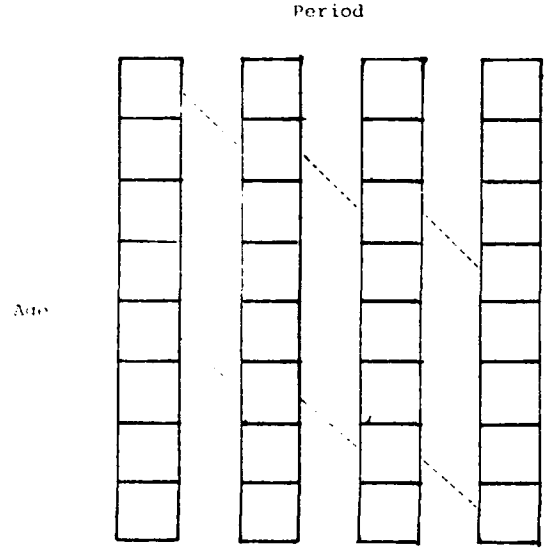


Figure 3. Age by Period Display of Repeated Cross-Sectional Surveys

underlies the partitioning. Hence, the accounting framework is not used optimally and the success of the identifying restriction employed is left in doubt.

A second problem arises in the estimation or calculation of effects depending on the design of the data structure, given that the data are multiple cross-sections. In particular, the data are inherently unbalanced. Although the data array is rectangular with respect to age and period, it is not rectangular with respect to cohort membership. Moreover, if the data structure were in some way augmented to make cohort membership balanced, the design would become unbalanced with respect to age and period. There is no solution to the problem, and it is a serious one, since it can influence the calculated effects to such an extent as to make them virtually useless.

A third problem is that data arrayed in the form of repeated cross-sections give stocks and not flows. That is, there is no information about cross-time linkages between individuals. An exception to this problem can occur if the response variable is irreversible and discrete. If the response is whether an individual "survives" from one period to the next, then we can determine cross-time linkages. The reason is straightforward. If we envision a cross-tabulation of "alive" vs. "dead" at time t against "alive" vs. "dead" at time t+1 then one of the cells must be zero--"dead" at time t and "alive" at time t+1. This fact, together with our knowledge of the univariate distributions

at each point in time, suffices to determine all cell frequencies in the four-fold table. This knowledge is of little use to us, however, if we seek to extend the problem by replacing the accounting categories with underlying variables and these variables are reversible. This is especially troublesome when we are dealing with population data that are in principle longitudinal but have been aggregated in a cross-sectional manner, thus destroying the cross-time links.

D. Retrospective Multiple-Cohort Studies Based on a Single Cross-Sectional Survey

This design locates individuals on the basis of a single cross-sectional survey, and elicits retrospective longitudinal data from them. Thus we get direct information from multiple cohorts for past periods. One way to structure such data is in the form of an upper-triangular period by cohort array, as in Figure 4. Here common age groups across cohorts can be linked by moving from upper left to lower right, provided that the period and cohort groupings are of the same size or length. Our information on the multiple cohorts then ends with the period of the cross-sectional survey, and there is full age-specific data for only the oldest cohort.

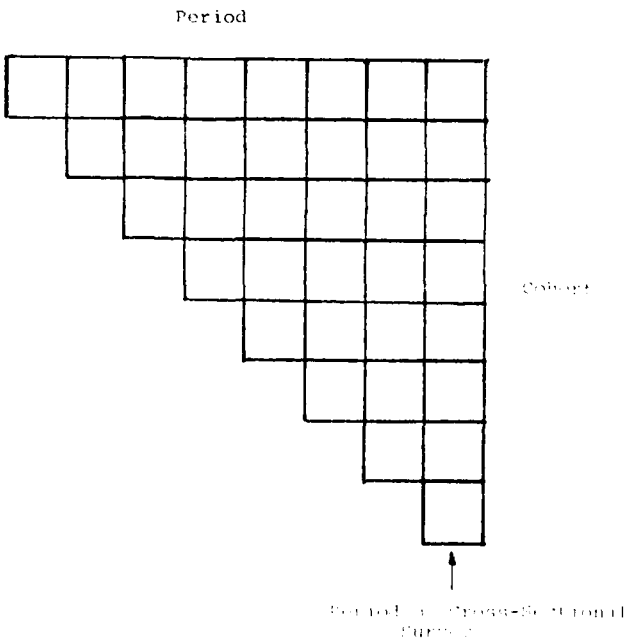


Figure 4. Retrospective Multiple-Cohort Study Based on a Single Cross-Sectional Survey

Birth histories ascertained from a single cross-section of marriage age women constitute a prime example of the application of this type of data structure. The World Fertility Survey, for example, collects this kind of information (International Statistical Institute, 1975).

The new feature the retrospective design presents is that there is complete preservation of the individual longitudinal records. Although this is a great asset, it is also a problem, since we must now deal with the dependence of data across time. Other problems created by this design include:

a) Memory decay.

b) Telescoping.

c) Selective reporting (particularly if the collection of data is official--of course in some societies the official status of a survey may mean to the respondent that the survey will brook no opposition).

d) Hidden attrition--this problem is of course present for any cross-section of cohorts. We are dealing only with the survivors of whatever the process may be. Or we may simply be dealing with survivors. Moreover, if we fashion a longitudinal APC accounting model there is still a linear dependency among age, period, and cohort with which we must deal.

E. Prospective Multiple Cohort Studies

We find these studies in psychology (Nesselroade and Baltes, 1974) and medicine, when there is a developmental process which is held potentially to vary with cohort. We begin with a cross-sectional sample, and group the individuals into cohorts. The key to this design is that we track all cohorts from a specified initial time period, with

subsequent follow-ups at regular intervals. We do not add new cohorts. The data have a lower-triangular cohort by period structure, as seen in Figure 5, which is an inverted image of Figure 4.

Data of this kind need not be longitudinal. In general, however, we would expect the data to be collected in panels or waves, thus permitting the study of "flows" as well as "stocks." This particular design is really nothing more than what is typically called a panel design or panel study. The emphasis we are placing on the design is unusual, however, since panel designs are usually implemented not because the investigators wish to study cohort processes, but because they are interested in "flows" as well as "stocks," i.e., because they are interested in "turnover."

One problem with this design is actual mortality (again the survivor problem), which will affect the older cohorts more strongly than it will the younger cohorts (except perhaps during a war period). A second problem is that this design is subject to panel attrition apart from differential mortality. Whether this attrition is related to cohort membership is unclear, but in general we would expect it to be. Finally, there is a design imbalance, so that effects of age, period and cohort in an accounting model have the usual dependency problems associated with estimated parameters in unbalanced models. The linear dependence or identification problem also remains.

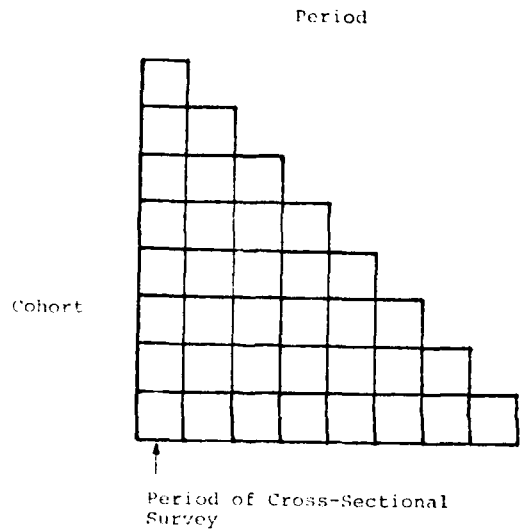


Figure 5. Prospective Multiple-Cohort Design

F. Staggered Prospective Multiple Cohort Studies

This design is similar to the preceding one except that each cohort is initiated at the same age, with subsequent follow-ups at regular intervals, as illustrated in Figure 6.

The advantage of this longitudinal design relative to the previous one is that all cohorts are observed the same number of times. Unfortunately this means that periods are unbalanced, and if their effects matter this design will not be optimal. The design suffers from the potential effects of panel attrition but not the effects of differential cohort mortality, unless something major happens to the environment (such as war or the discovery of the fountain of youth). The major disadvantage of this design is its cost in both time and money. As a consequence there are few good examples to point to. What is typically required is a major government data collection effort such as that associated with the National Assessment of Educational Progress sponsored by the U.S. Department of Education. Those data have thus far not been examined from the APC perspective.

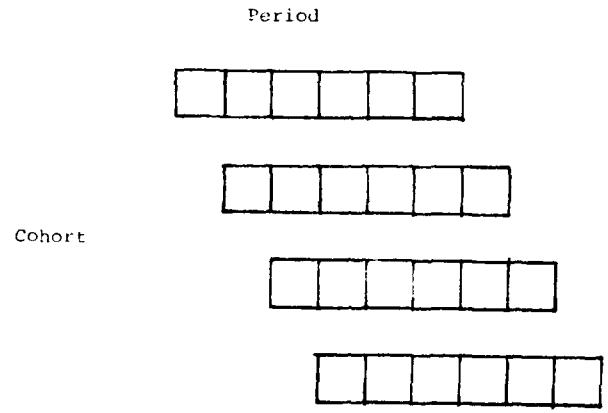


Figure 6. Staggered Prospective Multiple-Cohort Study

IV. APC ACCOUNTING MODELS

As used by a wide variety of investigators, the APC model is typically applied to a data array in the form of a cross-classification of age by period, or cohort by period or cohort by age. The same APC parameterization can be applied interchangeably to each of these three arrays--each containing the same information. For specificity we assume here the multiple cross-section design of Section III.C, and thus we are dealing with an age by period or I x J array, where the spacing of the I age categories is equal to the inter-period differences. Thus the K = I+J-1 diagonals of the array correspond to birth cohorts.

The basic APC model focusses on some parameter associated with a response variable, Y, e.g.,

$$\theta = \mathcal{E}(Y) \tag{2}$$

where $\mathcal{E}$ denotes the expectation operator and Y is treated as a continuous random variable, or

$$\theta = \log(p/(1 - p)) \tag{3}$$

where p = Pr(Y = 1) if Y is a dichotomous random variable. The model then expresses θ as a linear function of age, period, and cohort effects (α's for ages, π's for periods, γ's for cohorts), i.e.,

$$\theta_{ijk} = f(\alpha_i, \pi_j, \gamma_k) = \mu + \alpha_i + \pi_j + \gamma_k \tag{4}$$

subject to the restrictions

(5) $\sum_{i=1}^{i=I} \alpha_i = \sum_{j=1}^{j=J} \gamma_j = \sum_{k=1}^{k=K} \gamma_k = 0$,

with K = I+J-1. The k-subscript, which indexes cohorts, is of course redundant but we include it here in the specification to remind us of the symmetry of age, period, and cohort in the model. The linear restrictions in (5) represent an arbitrary choice on our part, and any other of the standard ANOVA-like conventions, such as setting the effects of the first levels of the variables equal to zero, would suffice. When the APC model was first introduced investigators realized that the model could not be made more complex by the introduction of unrestricted interaction terms of the form $(\alpha\gamma)_{ik}$ given the terms already present in (4). Yet much of the discussion in the cohort literature focusses qualitatively on just such interactions, e.g., cohort-specific age effects. It is possible to introduce restricted forms of interaction, and we do so in Section VI.

To complete the specification of the APC model of expressions (4) and (5) we need to do something about the linear components of the effects. Since

Cohort = Period - Age

the linear component of any one set of effects is either the sum or the difference of the linear components of the other two sets of effects. This is known as the APC "identification problem." Given the specification of (4)

and (5), the nonlinear components of the effects of age, period and cohort are identifiable--even if the linear components are not.

When the specification of expression (4), developed here for use with multiple cross-sections, is used for longitudinal studies, the same identification problems remain. That is, panel data and multiple cross-sections are equally informative about the parameters of the APC accounting model. But since panel data allow the modelling of flows as well as stocks, investigators working with such data should develop a richer class of models than the APC accounting models discussed here.

It has been suggested that the crudeness of the age and cohort classification scheme used in the multiple cross-section I x J array leads to the APC identification problem (Winsborough, 1976). There are two ways to think about this suggestion. First, we can think of one of the classifications in the data array being made more refined while leaving the other classification as is. For example, we might envision use of a finer age gradient, while keeping the period gradient constant and less refined. Fienberg and Mason (1978) consider this case, and point out that not only does the original identification problem remain, an additional one is created. If it is supposed that the period classification is made more refined, doing nothing to the age classification, it is possible that the original identification problem is solved, but this can only be due

to a loss in ability to track individuals through time correctly. Thus, either the identification problem is not solved, or it is solved at the expense of creating an even bigger problem.

A second way to use the suggestion is to think in terms of continuous age, period, and cohort measurements. We do this below. Of course, cohort membership defined as an instantaneous event runs counter to the typical social science use of the term. We entertain this definition only to explore the implication of continuity.

Reformulating expression (4) in terms of continuous age (A), period (P), and cohort (C) gives

$$\theta_{APC} = f(A, P, C) \tag{6}$$

where

$$C = P - A \tag{7}$$

and f is taken as a polynomial in A, P, and C. Thus, for example, the 2nd-order or "response surface" APC model becomes

$$\theta_{APC} = \beta_0 + \beta_1 A + \beta_2 P + \beta_3 C + \beta_{11} A^2 + \beta_{12} AP + \beta_{22} P^2 + \beta_{13} AC + \beta_{23} PC + \beta_{33} C^2 \tag{8}$$

Here the β's are regression-like effect parameters.

Does this reformulation solve the identification problem? The answer is no, because the basic linear identification problem, associated with expression (7), remains. Indeed, we now must ask if the remaining 2nd-order effects are identifiable. Using expression (7) we have that

$$C^2 = P^2 + A^2 - 2PA \tag{9}$$

$$CP = P^2 - AP \tag{10}$$

$$CA = PA - A^2 \quad . \tag{11}$$

Thus, only three of the six 2nd-order coefficients can be identified, and if we choose to resolve this identification problem by focussing on the effects of A^2, P^2, and C^2 we can not pick up interaction effects at the 2nd-order level. Alternatively, eliminating the squared terms will allow estimation of the 2nd-order interactions, but the two alternatives are equivalent. Thus, the choice between

$$\theta_{APC} = \beta_0 + \beta_1 A + \beta_2 P + \beta_3 C + \beta_{11} A^2 + \beta_{22} P^2 + \beta_{33} C^2 \tag{12}$$

and

$$\theta_{APC} = \beta_0 + \beta_1 A + \beta_2 P + \beta_3 C + \beta_{12} AP + \beta_{13} AC + \beta_{23} PC \tag{13}$$

must depend on information external to the analysis.

There is a related point here, which has more, perhaps, to do with parsimony than it does with questions of identification. Suppose we use the following model

$$(14) \qquad \theta_{APC} = \beta_0 + \beta_1 A + \beta_2 P + \beta_3 C + \beta_{11} A^2 + \beta_{22} P^2 + \beta_{33} C^2 \ .$$

This model reflects an initial choice to give precedence to C^2, P^2 and A^2 over AC, AP and CP. Use of the model does not necessarily mean that we think the linear by linear interactions are unimportant, but could as well reflect our triage system for the identification problem. With a bit of luck, we might discover upon estimation that the coefficients for $-C^2$, P^2 and A^2 were essentially equal. Then expression (9) shows that we could replace (14) by

$$(15) \qquad \theta_{APC} = \beta_0 + \beta_1 A + \beta_2 P + \beta_3 C + \beta_{12} AP \ ,$$

which has a linear by linear interaction term. Model (15) is simpler than model (14), has been arrived at from a consideration of parsimony, and shifts the interpretation from a purely "additive" to an interactive one by recasting the meaning of nonlinearity in this instance. Here, as elsewhere, the more parsimonious representation is helpful to the extent results obtained with it are interpretable. We point out the potential for translation from (14) to (15) because even here, where we place so much stress on the use of external information to make decisions concerning

identifying restrictions, it is still worthwhile to bear in mind the possible gains in clarity that explicit concern for parsimony can yield.

When we choose $\theta(t)$ to be an m<u>th</u>-order polynomial in A, P, and C, at the m<u>th</u>-level there are

$$\binom{m+2}{m} = (m+2)(m+1)/2$$

regression coefficients, of which m+1 are estimable. Only three of these are powers: A^m, P^m and C^m. For this reason m-2 m<u>th</u>-order interaction terms will be estimable as well (for $m > 2$). Thus, while the polynomial approach does not eliminate the identification problem, it does suggest a way to get at higher-order interactions. We discuss this approach further in Section VI.

Finally, continuing with the polynomial approach, when $\theta(t)$ is the <u>sum</u> of an (I-1)<u>th</u>-order polynomial in A, a (J-1)<u>th</u>-order polynomial in P, and a (K-1)<u>th</u>-order polynomial in C, i.e.,

$$(16) \qquad \theta_{APC} = f_A(A) + f_P(P) + f_C(C) \ ,$$

we can use expression (16) interchangeably with expression (4) in analyzing the usual multiple cross-section data array (Winsborough, 1976).

V. EXTENDING THE APC MODEL TO INCLUDE INTERACTIONS

An implicit message in Ryder's (1965) distinguished paper, and an explicit message in Glenn's (1976) discussion, is that for many, perhaps most, problems it is desirable to include age-cohort or other interactions. Fienberg and Mason (1978) touch on this matter and conclude that inclusion of such interactions will require even more identifying restrictions (see also the discussion in Section IV). Although virtually all of the APC modelling to date has assumed that the inclusion of interaction terms is not possible, and our earlier paper (Fienberg and Mason, 1978) suggests the same, it is possible to extend the basic APC model to include interactions.

To understand why such a possibility exists, we continue considering an I x J, age by period array. Were we to fit an additive age-period model there would be (I-1)(J-1) degrees of freedom (d.f.) for assessing the fit of the model. As noted by Fienberg and Mason (1978), the inclusion of a set of cohort effects in this kind of model is a way to get a simple and parsimonious description of age by period interactions. For the APC model of expression (4) there are an additional I+J-2 cohort parameters (taking into account the linear constraint in (5)), and one unidentifiable linear effect, leaving

$$(I - 1)(J - 1) - (I + J - 2) - 1 = (I - 2)(J - 2)$$

d.f. for assessing fit.[3] Clearly, then, there must be room for modelling age by period interaction effects over and above the effects captured by cohorts. Similarly, if we view period effects as simple and parsimonious descriptions of age by cohort interaction effects, then the (I-2)(J-2) residual d.f. must leave room for the modelling of additional age by cohort effects over and above the effects captured by period.

To date we have conceived of four different strategies for extending the APC model to include additional interaction effects. Three of these are suggested in part by the polynomial version of the APC model discussed in Section IV, while the fourth is in the spirit of the examination of residuals for a small number of outliers.

A. Polynomial Models

Recall from Section IV that for continuous age, period, and cohort variables our model takes the form of expression (6) subject to the restriction of expression (7). In addition, imposing the linear restriction embodied in expression (16), there are three 2nd-order polynomial terms (A^2, P^2, and C^2), and we have noted that their inclusion prevents us from including in an identifiable way the other three 2nd-order terms (AP, AC, and PC).

In general, for the mth-order model there are m+1 identifiable terms at the mth level, and the model in expression (16) uses only three of these. Thus, at the mth-order there are potentially m-2 d.f. available for modelling interaction terms.

Suppose we wish to model age by cohort interactions. At the mth-order there are m-1 such terms:

$$AC^{m-1},\ A^2C^{m-2},\ \ldots,\ A^{m-1}C \quad .$$

Since

$$(A + C)^m = P^m \quad ,$$

only m-2 of these will be identifiable, and it is in a sense arbitrary which we think of ourselves as excluding. For the present discussion, we adopt the convention that the coefficient of $A^{m-1}C$ equals zero.[4] Thus, we can use up the residual m-2 d.f. at the mth level by adding in all age by cohort interaction terms at that level, subject to an identifying restriction.

To adopt this approach in practice we could proceed hierarchically, first adding in age by cohort terms of the 3rd order, then the terms of the 4th order, etc. Thus we would replace (16) by

$$\theta_{APC} = f_A(A) + f_P(P) + f_C(C) + f_m(A,C) \qquad (17)$$

where $f_m(A,C)$ is a polynomial which includes all age by cohort interaction terms up to the mth-order. For m = 3 this model uses m-2 = 1 extra d.f., and thus I ≥ 4 is required for there to be any residual d.f. for assessing the fit of the model, and for a 3rd order interaction model to have meaning. Similarly, for m = 4 the model uses 3 extra d.f. and we require that I ≥ 5, and J ≥ 4 if I = 5, to assess the fit of the model.

A second approach to fitting such polynomial interaction terms for age by cohort is to single out those interaction terms that involve a linear component of age, i.e.,

$$AC^2,\ AC^3,\ \ldots,\ AC^{K-1} \quad .$$

Thus, instead of (17) we employ

$$\theta_{APC} = f_A(A) + f_P(P) + f_C(C) + \beta_{133}AC^2 + \beta_{1333}AC^3 + \ldots + \beta_{13\ldots3}AC^{K-1} \quad . \qquad (18)$$

This model uses up the extra d.f. at the 3rd-order level, and one extra d.f. at each order up to K-1. The d.f. associated with (18) are

$$(I - 2)(J - 2) - (I + J - 3) = (I - 3)(J - 3) - 2 \quad ,$$

and we require that min(I,J) ≥ 4.

To the model of expression (18) we might also add a set of age by period interaction terms that involve a linear component of age

(19) $\beta_{122}AP^2 + \beta_{1222}AP^3 + \ldots + \beta_{12\ldots2}AP^{J-1}$,

which would use up an extra J-2 d.f., leaving

$$(I - 4)(J - 3) - 3$$

d.f. for assessing goodness-of-fit. To do this requires $I \geq 5$ and $J \geq 4$. In order to be successful in this approach, leaving enough d.f. to assess the fit of our model, we need a relatively large data array.

B. Polynomial Models with Induced Metrics

One way to take advantage of the models introduced in the preceding subsection without using too many d.f., is to introduce a metric or set of numerical scores for one of the accounting dimensions. Suppose for example that, on the basis of prior or external information, the score v_i is assigned to the ith category of age. Then we can replace $f_A(A)$ in (16) by v_i for the ith category of age, and use

(20) $\beta_{13}v_iC + \beta_{133}v_iC^2 + \beta_{1333}v_iC^3 + \ldots + \beta_{13\ldots3}v_iC^{K-1}$

and

(21) $\beta_{122}v_iP^2 + \beta_{1222}v_iP^3 + \ldots + \beta_{12\ldots2}v_iP^{J-1}$

for the age by cohort and age by period interaction components. This model is in the spirit of that given by expression (18) with (19) added to the right-hand side. It finesses the linear identification problem by inducing a nonlinear metric on age, and gains us I-2 d.f. By dropping

the quadratic term in age, however, we have open 1 d.f. at the 2nd-order, and thus we have added the term $\beta_{13}v_iC$ to the front end of expression (20) and used an extra d.f. The d.f. for assessing goodness-of-fit have therefore increased by I-3, to

$$(I - 4)(J - 3) - 3 + (I - 3) = (I - 4)(J - 2) - 2 \quad .$$

The critical feature of such an approach is that external information must be used to develop the induced metric.

Johnson (1981) employs the approach just outlined in an analysis of age-specific marital fertility schedules. His induced metric for age rests heavily on "the assumption that marital fertility has a regular age pattern in all human populations," and he derives a set of $\{v_i\}$ based on the work of Coale and Trussell (1974). We give Johnson's (1981) values for $\{v_i\}$ in Table 2, together with his data, which consist of counts of the observed number of births (b_{ijk}) and the total number of expected births under natural fertility (n_{ijk}) for I = 12 two-year age groups and J = 4 two-year periods.

Johnson assumes that the $\{b_{ijk}\}$ are realized values of independent Poisson random variables with means $\{m_{ijk}\}$. He models

(22) $\theta_{ijk} = \log(m_{ijk}/n_{ijk})$

Table 2. Induced metric for age categories, and age by period display of observed and expected births under natural fertility (in parentheses) from Davao, Philippines sample surveys as given by Johnson (1981).

Age Group	Age Metric(v_i)	Period 1971-71	1973-74	1975-76	1977-78
20-21	-.02	$269_{(301)}$	$261_{(288)}$	$151_{(202)}$	$79_{(100)}$
22-23	-.08	$405_{(471)}$	$329_{(383)}$	$218_{(286)}$	$157_{(178)}$
24-25	-.13	$411_{(490)}$	$389_{(514)}$	$282_{(330)}$	$167_{(236)}$
26-27	-.28	$342_{(430)}$	$302_{(466)}$	$312_{(406)}$	$161_{(236)}$
28-29	-.41	$328_{(462)}$	$261_{(397)}$	$186_{(334)}$	$171_{(283)}$
30-31	-.56	$305_{(452)}$	$217_{(405)}$	$133_{(276)}$	$120_{(223)}$
32-33	-.72	$254_{(435)}$	$214_{(385)}$	$145_{(287)}$	$91_{(184)}$
34-35	-.86	$224_{(383)}$	$180_{(376)}$	$113_{(264)}$	$66_{(186)}$
36-37	-1.00	$163_{(271)}$	$142_{(311)}$	$124_{(252)}$	$52_{(173)}$
38-39	-1.14	$104_{(228)}$	$98_{(297)}$	$78_{(195)}$	$59_{(159)}$
40-41	-1.28	$80_{(171)}$	$75_{(158)}$	$36_{(111)}$	$31_{(103)}$
42-43	-1.42	$40_{(100)}$	$32_{(98)}$	$23_{(74)}$	$17_{(51)}$

where the specification for the $\{\theta_{ijk}\}$ is as described above. If we take the $\{n_{ijk}\}$ as fixed (known) positive scale factors, then expression (22) can be written as a loglinear model for the $\{m_{ijk}\}$,

$$(23) \qquad \log m_{ijk} = \log n_{ijk} + \theta_{ijk}$$

where the $\{n_{ijk}\}$ are used as "offsets" or initial values in an iterative maximum likelihood calculation (Haberman, 1978:124-133).

Since I = 12 and J = 4 for the data in Table 2, the final model of the preceding subsection has only 5 d.f., whereas the model just described with added interaction terms given by expressions (20) and (21) has 14 d.f. This model fits the data well (G^2 = 17.3 with 14 d.f.), as does the basic APC model of expression (14), with $f_A(A)$ replaced by αv_i for the i<u>th</u> category of age (G^2 = 33.1 with 29 d.f.). Comparing likelihood ratio statistics we find

$$\Delta G^2 = 33.1 - 17.3 = 15.8$$

with 15 d.f., and thus have a basis for settling on the original APC specification as modified by the induced metric on age. In fact, the age-cohort model (with no period effects) fits the data extremely well.

C. <u>Tukey 1 d.f. Models and Their Generalizations</u>

The approaches for the inclusion of interactions described up to this point are based on 1 d.f. interaction components added to the basic APC model, and depend on

interactions defined explicitly by nonlinear combinations of age, period or cohort. It is also possible to define 1 d.f. interactions which structure main effects. Models whose interaction components amount to a structuring of main effects have their origins in a test for interaction devised by Tukey (1949) for two-way ANOVA without replicates.

Within the APC framework, a natural analogue to the Tukey 1 d.f. model might be, for example, one which structures interactions between age and cohort in terms of the age and cohort main effects, in the following way:

$$\theta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \delta\alpha_i\gamma_k \quad . \tag{24}$$

This model introduces a single new interaction parameter, δ, for interactions between age and cohort, and thus has (I-2)(J-2)-1 d.f.

Expression (24) can be thought of as arising from latent constructs for age and cohort. Suppose there are unmeasured variables U(A) and U(C) associated with age and cohort effects, respectively. Further suppose that U(A) ≠ A and U(C) ≠ C (otherwise there is no point to this formulation), where A and C are given their usual scoring in sequentia. (A) and calendar (C) time. If we actually had measurements on U(A) and U(C) we could fit

$$\theta_{ijk} = \mu + \alpha U(A) + \gamma U(C) + \beta_j + \lambda U(A)U(C) \quad . \tag{25}$$

This would amount to an instance of the polynomial approach with induced metric, discussed earlier. Substituting α's for $\eta U(A)$ and γ's for $\xi U(C)$ in (25) yields

$$\theta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \lambda(\alpha_i\gamma_k)/\eta\xi \tag{26}$$

where

$$\delta = \lambda/\eta\xi$$

so that δ differs from λ only by arbitrary scale factors for U(A) and U(C). Thus δ can be interpreted as the linear effect of the multiplicative term involving the two unmeasured age and cohort constructs.

The notion of cohort trajectories provides another way of understanding the model given by expression (24). Following a suggestion by O. D. Duncan, suppose we look at the quantities

$$\omega_{ijk} = \theta_{ijk} - \theta_{i-1,j-1,k} \tag{27}$$

for $2 \le i \le I$ for the kth cohort. These quantities describe the successive first-differences or trajectory for this cohort. Under the basic APC model of expression (4), we get

$$\omega_{ijk} = (\alpha_i - \alpha_{i-1}) + (\beta_j - \beta_{j-1}) \tag{28}$$

$$= \omega_i + \omega_j \quad ,$$

where the Δ's denote first-differences. For this simplest case, the description of trajectory is independent of cohort, as it must be given the "additive" nature of the model. When we substitute in (27) for the model given by expression (24), however, we get

$$(29) \qquad \omega_{ijk} = \Delta\alpha_i + \Delta\pi_j + \delta(\Delta\alpha_i)\gamma_k \quad .$$

In (29) the effects of cohort on trajectory enter into the model for the $\{\omega_{ijk}\}$ explicitly in the form of a reduced-parameter version of the interaction between age and cohort. That is, the contribution of age to the trajectory of each cohort differs in a constrained way across cohorts.

On first encounter the Tukey 1 d.f. approach may appear artificial, perhaps because of the rationale we have offered in terms of a linkage between main effects and *unmeasured constructs*. This is probably a consequence of unfamiliarity. Although we would not claim the universal applicability of the Tukey 1 d.f. approach, the linkage between main effect and unmeasured variable may be just what the analyst wants in situations where use of an APC model is contemplated. In these cases, the analyst typically does not have available measures of the variables presumed to underlie age, period or cohort differences. In addition, variables which are merely linear transformations of scaled A, P or C provide little analytic leverage. But the analyst may well have an unmeasured process in mind which is reflected in a hypothesized pattern of *main effects*. Thus,

there are undoubtedly situations in which the extension of the Tukey 1 d.f. approach to the APC modelling context may be quite suitable.

We could go further than the 1 d.f. interaction model of expression (24), by assuming a multiplicative age by cohort interaction where the age and cohort parameters in the interaction are different from those in the basic model, i.e.,

$$(30) \qquad \theta_{ijk} = \mu + \alpha_i + \pi_j + \gamma_k + \phi_i\gamma_k \quad .$$

In this instance we use the data to induce metrics on the ordered age and cohort categories, in order to allow the variables so scaled to interact multiplicatively. This approach uses up 2I+J-4 extra d.f. (not just 1 d.f.) and the model thus has (I-3)(J-4)-12 d.f. Fitting such a model requires a truly large data array. For example, we could not apply the model to the data of Table 2, even if we had an additional period.

For counted data where θ_{ijk} represents the log expected value as in expression (23) or the log-odds or logit as in expression (3), Chuang (1980) describes a modification of the Newton-Raphson procedure that can be used to estimate parameter values for these models.[5]

D. Using Dummy Variables to Adjust for Outliers

In assessing the fit of the APC model to a set of data we want to look at not only summary measures but also standardized residuals, say of the form

$$(31) \qquad r_{ijk} = \frac{Y_{ijk} - \hat{\theta}_{ijk}}{\text{estimated S.E. } (Y_{ijk} - \hat{\theta}_{ijk})} \quad .$$

In some cases, a poor fit of the APC model will be reflected by large standardized residuals (either positive or negative) for several cells. In others, one or two residuals will stand out.

Suppose the residual for cell (i,j) is very large, and thus is thought of as an "outlier." Then we could set that cell aside and fit the APC model to the remaining ones. This is equivalent to using up 1 d.f. for a dummy variable associated with cell (i,j), and is yet another way to introduce interaction into the basic model.

The dummy variable approach for outliers can be used to deal with more than one discrepant cell either by introducing a dummy variable for each outlier, by using a single dummy variable for several outliers whose residuals from the APC model are of similar size or by using other criteria for grouping cells. Mason and Smith (this volume) use external information to define an indicator variable, closely related to a dummy, whose inclusion substantially improves the fit of their model in an analysis of tuberculosis mortality. The outliers in this instance reflect the government's shifting of healthy civilian men into the armed forces during World War II while screening tubercular cases from entry into the military, which led to

artifactual increases in tuberculosis mortality for younger men in the civilian population during the 1940s. Mason and Smith deal with this phenomenon by constructing an age-period interaction term which is one outside the affected age-period groups, and otherwise takes on values reflecting the decrease in the civilian population base for each affected age group during World War II. Thus, these authors use a single degree of freedom to draw into the model a "nuisance" phenomenon which cuts across several outlier cells. The justification for a procedure such as this is enhanced by a priori substantive understanding of the problem, as usual.

In assessing the fit of the basic APC model we have noted previously that the fitted values $\{\theta_{ijk}\}$ remain the same no matter how we choose to resolve the linear identification problem, provided that we do not resort to over-identification. The standardized residuals are likewise unaffected by the resolution of the identification problem. Indeed, no linear restriction whatsoever is required to obtain fitted values and residuals (Fienberg and Mason, 1978:18).

VI. DISCUSSION

We've ranged from broad gauged discussion of perspectives to the narrowly statistical. What have we accomplished?

To begin with, we have attempted to provide a framework in which to think about the various purposes to which APC modelling and quantitative cohort analysis in general can be put. Our conclusion is that quantitative cohort analysis can be used for a variety of purposes, including the extremes of descriptions embedded in historical settings and inferences about social mechanisms thought to exert their forces indefinitely.

Second, we have suggested that although cohorts are often thought of as groups, a cohort perspective ideally involves two or more levels of conceptualization to flesh out the framework. An empirical realization of a multi-layered framework is difficult to achieve, but remains a worthy goal for the conceptual clarity it can promote.

Third, we have shown that, given a multi-level framework in which connections are established between macro forces and individual actors, it is possible to assess the impact of data aggregation in terms of coefficient bias. In general, the commonly intuited connotation of "grouped" data is misleading. There is not necessarily a direct mapping between the conceptual units of analysis and the refinement inherent in a particular degree of data aggregation. A

best aggregation amounts to a compact data storage and presentation device; at worst a roadblock which can not be bypassed.

Fourth, different kinds of data structures can be used for different cohort analytic approaches. We have reviewed and characterized these structures.

Fifth, we have shown that the identification problem in APC models is not due to crudeness of measurement. It remains even when measurement of age, period and cohort is infinitely refined in calendar and sequential time.

Lastly, we have described four ways to extend APC models to include interactions beyond those represented by one of the accounting dimensions, given the presence of the other two.

We were led to formulate a view on the scope of quantitative cohort analysis because few cohort analysts make their own stance clear in the course of their research. This can only add to the difficulties in rendering judgments as to the appropriateness and adequacy of the model chosen. Few critics of cohort analyses have clear ideas of the intellectual needs which motivate their assessments, and can hardly be expected to supply goals left unstated in the work they examine. We hope that our views on scope and purpose will be found workable by others, or revised until they become helpful, so that particular empirical quantitative cohort analytic studies can be both carried out and evaluated productively.

We introduced the idea of multi-level analysis into the discussion for two reasons. First, it is well known that a goodly supply of data, when combined with access to a computer, creates an almost irresistible urge to carry out statistical manipulations. Many such urges have been satisfied by fitting APC models. Doubtless every statistical technique or approach receives its share of this kind of devotion. We have attempted to point to a minimum requirement for a conceptual framework to be used in association with an APC model, or, for that matter, any quantitative cohort model. The specification of a multi-level model is a formidable requirement; if followed it should offset the ease with which APC specifications can be identified, and it should lead to worthwhile substantive contributions. Koopmans' (1949) discussion of the need for substantive theory in connection with identifying restrictions in simultaneous equations models is equally applicable in the cohort, and specifically APC, modelling context.

A second reason for introducing the idea of multi-level analysis into the discussion of APC modelling is that there has been for some time confusion about the conceptual units of analysis in cohort studies, and the problem has been exacerbated by an awareness of data aggregation. We have exploited a two-level conception of variables, as well as a classification of types of explanatory variables, to trace the implications of data aggregation. The results are

too detailed to recapitulate here, but we hope that a study of them will at least banish the mistaken association often made between "grouped" data and "group" models.

There were two reasons for discussing identification in "additive" APC models yet again. First, we wanted to show that an argument sometimes advanced in defense of APC specifications--that the identification "problem" would go away if only we had infinitely refined measurement of age, period and cohort, is just not correct. APC modelling must therefore be defended on other grounds. Second, it turned out that the line of attack used on the continuous variable representation of age, period and cohort was a suggestive one for the study of interactions above and beyond the "main effects" of age, period and cohort.

Finally, we presented the material on additional interactions in APC models because of the demand for the extension of the APC framework to deal with phenomena whose conceptualization may require more than "main effects" for age, period and cohort. The availability of a variety of approaches to interaction in the APC accounting framework disarms a major criticism, and could materially enhance the framework's usefulness.

Apart from the omission of interactions there have been a number of criticisms of APC modelling, and it may be useful to review them in light of the foregoing discussion.

One objection to APC specifications holds that the need to make an identifying restriction is evidence of the impossibility of the task. According to this view, age, period and cohort effects can not be separated because one of the accounting dimensions is a constrained interaction of the other two. As is well known, the variable which "carries" the interaction in a generalized linear model expression can not vary independently of its constituents. Thus, according to the argument, it makes no sense to construct a model in which age, period and cohort are thought to play separate explanatory roles, because no single dimension can vary by itself.

This view seems to assume that the user of an APC specification has no substantive hypotheses or theoretical model concerning the relevance or salience of cohort membership. It is literally true that cohort membership can not vary independently of age and period. In fact, the generalization of this point holds for interactions of any sort in any analysis of variance problem. One way of dealing with interaction, suited to many problems, is to view the response variable as a nonadditive combination of the explanatory variables. Another way to deal with interaction is to try to conceptualize the phenomenon in terms of other variables. This alternative can hardly be said to be inferior to the former approach. It is identical to the approach which APC specifications allow and are consistent with. Following this approach, one begins with

conceptualization and attempts to move to explicit measurement, to test understanding of the interaction. This is always a goal in APC specifications, if not always an attainable one. Insofar as the analyst does not reify the accounting categories, that is, does not make the mistake of thinking that each of the accounting dimensions is indistinguishable from its theoretical meaning for a specific substantive phenomenon, there is no logical impediment to APC modelling. It is true that some empirical analysts have failed to specify their substantive framework, but such errors should not be confused with the properties of the accounting framework itself.

A second objection to APC specifications holds that the "additive" APC model requires the analyst to <u>believe</u> that cohort contrasts, say, are constant over ages and periods. This is incorrect. There is a difference between attempting to work with an approximation of something more complex, and believing that the reality is actually the approximation. Nothing in the additive APC framework prevents the analyst from supposing that along the diagonals of an age by period array the interaction (i.e., cohort) coefficients vary. Rather, it suffices to suppose that whatever the within cohort variability, the differences between cohorts are meaningful. This is again a common stance in analysis of variance. Rare is the data set--and much cherished for classroom illustrations--which exhibits perfect additivity. As long as the departures from

additivity satisfy certain conditions (e.g., not "too" large, and no pattern to them), we are usually satisfied with the approximation provided by the additive model in the typical r x c case. Likewise, when we work with an interaction model for a multiway problem, we often choose to ignore certain higher-order interactions. This is the same point all over again: We are working with an approximation.

A third objection to APC specifications, touched on above, contends not only that they are inherently additive, but also that additive APC specifications are generally uninteresting, or at most of interest in particular substantive areas (e.g., demography). Regrettably, we have in the past contributed to the view that APC models are inherently additive, but have taken the opportunity here to clarify the possibilities and provide new results which show that it is not necessary to restrict one's attention to purely additive APC models. Whether additive APC specifications are of limited use, or of use in certain fields but not in others, is of course separate from the presumed unavailability of interactions in the accounting framework. The claims of limitation are certainly untested empirically. Moreover, given the depth of substantive reasoning we espouse for the application of the accounting framework, we do not soon expect to see a compelling disquisition showing that additive APC specifications are generally less helpful than interactive ones, or that

certain substantive areas are better suited than others to additivity. The connections between theory and research seem too complex to validate such conclusions.

A fourth objection to APC specifications is that they are atheoretical and should be eschewed in favor of models which replace the accounting dimensions with measured variables. The charge that APC specifications are atheoretical is incorrect, as we have already discussed. Particular analyses may be atheoretical, but that is not a function of the framework. Moreover, the use of measured variables is hardly a guarantee of the presence of theory. This objection is also unreasonable in its insistence on the use of measured variables when there are clearly so many instances where measures of any sort are seemingly impossible to find. Nevertheless, as we have said before, the use of measured variables is an important goal for any research involving quantitative cohort analysis. The use of available information is always an appropriate goal.

We have tried to clear away the underbrush of misconceptions which has surrounded the use of APC specifications. For those who would use them we urge attention to the caution that considerable substantive reasoning must be brought to bear <u>prior</u> to data analysis. For those who would argue that the accounting framework is inherently meaningless, we have tried to show that the arguments we are aware of are invalid. APC specifications can not be applied willy-nilly, but they can be used to

adjudicate between compatible reduced models, to estimate full specifications when there is justification, and particularly when indicators of the substantive phenomena which underlie the accounting dimensions are lacking. Any quantitative cohort analysis is a form of time-series analysis. As such, the development and estimation of a cohort model requires sensitivity and delicacy; this has been underlined for us by the empirical work of Mason and Smith (this volume). The APC accounting framework is not something one needs to be "for" or "against." It is, rather, a framework that can be used when alternatives are unavailable or less appealing.

FOOTNOTES

[1]If the competing specifications are compatible, then an evaluation is possible using an age-period-cohort parameterization. If the alternatives are incompatible, then their formal evaluation is carried out using procedures developed for the evaluation of nonnested models (Cox, 1961; Atkinson, 1970; Quandt, 1974). Since theoretical reasoning surrounding the use of age, period and cohort in models of social processes has thus far typically not been so precise as to produce many instances of competing nonnested reduced models, it appears that the class of applications we direct our attention to contains perhaps the majority of cohort analyses.

[2]It is also the case that micro phenomena can have an impact at the macro level. The demonstration of this requires temporal data. Mason (1980) focusses on situations in which the only available data are cross-sectional, and does not treat causation from the micro to the macro level. The Easterlin (1961) argument for the generation of population waves (see also Smith, 1980) hypothesizes a micro process for boom and bust fluctuations in fertility which goes well beyond the purely mechanical.

[3]For the basic APC model we require that $\min(I,J) \geq 3$ for there to be residual d.f. with which to assess the fit of the model.

[4]This is arbitrary, and similar to fixing $\beta_3 = 0$ in equation (8) to resolve the linear identification problem. Clearly, the resulting parameter estimates make sense only if this identification choice is grounded in substance. On the other hand, our estimates of the Θ's will be the same no matter which convention we adopt to deal with this new identification problem at the mth order, and we can use an arbitrary convention to get an indication of the importance of the mth-order age by cohort interaction terms.

[5]For extensions, generalization and discussions of the Tukey 1 d.f. model in conventional ANOVA see Johnson and Graybill (1972), Mandel (1969, 1971), and Cook (1975).

REFERENCES

Atkinson, A. C.
1970 "A method for discriminating between models." Journal of the Royal Statistical Society, Series B 32:323-353.

Brown, B. W., Jr.
1978 "Statistics, scientific method and smoking." Pp. 59-70 in J. M. Tanur et al., Statistics: A Guide to the Unknown, Second Edition. San Francisco: Holden-Day.

Carr-Hill, R. A., K. Hope, and N. Stern
1972 "Delinquent generations revisited." Quality and Quantity 6:327-352.

Chuang, J-L. C.
1980 Analysis of categorical data with ordered categories. Ph. D. Dissertation. Minneapolis-St. Paul: School of Statistics, University of Minnesota.

Coale, A. J. and T. J. Trussell
1974 "Model fertility schedules: variations in the age structure of childbearing in human populations." Population Index 40:185-257.

Cook, R. D.
1975 "Analysis of interactions." Unpublished manuscript, Minneapolis-St. Paul: School of Statistics, *University of Minnesota.*

Cox, D. R.
1961 "Tests of separate families of hypotheses." Pp. 105-123 in Vol. 1 of J. Neyman (ed.), Fourth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, California: University of California Press.

Easterlin, R.
1961 "The American baby boom in historical perspective." American Economic Review 51:869-911.

1978 "What will 1984 be like? Socioeconomic implications of recent twists in age structure." Demography 15:397-432.

Farkas, G.
1977 "Cohort, age and period effects upon the employment of white females: evidence for 1957-1968." Demography 14:843-861.

Featherman, D. L. and R. M. Hauser
1975 "Design for a replicate study of social mobility in the United States." Pp. 219-251 in K. C. Land and S. Spilerman (eds.), Social Indicator Models. New York: Russell Sage Foundation.

Fienberg, S. E., and W. M. Mason
1978 "Identification and estimation of age-period-cohort models in the analysis of discrete archival data." Pp. 1-67 in K. F. Schuessler (ed.), Sociological Methodology 1979. San Francisco: Jossey-Bass.

Glenn, N.D.
1976 "Cohort analysts' futile quest: statistical attempts to separate age, period and cohort effects." American Sociological Review 41:900-904.

Greenberg, B. G., J. J. Wright, and C. G. Sheps
1950 "A technique for analyzing some factors affecting the incidence of syphilis." Journal of the American Statistical Association 251:373-399.

Haberman, S. J.
1978 Analysis of Qualitative Data. Vol. 1: Introductory Topics. New York: Academic Press.

Hartley, H. O., and R. L. Sielken
1975 "A 'super-population viewpoint' for finite population sampling." Biometrics 31:411-422.

International Statistical Institute
1975 The World Fertility Survey: The First Three Years, January 1972--January 1975. The Hague-Voorburg, Netherlands: International Statistical Institute, 428 Prinses Beatrixlaan.

Johnson, E. E. and F. A. Graybill
1972 "Analysis of a two-way model with interaction and no replication." Journal of the American Statistical Association 67:862-868.

Johnson, R. A.
1981 "Analysis of age, period and cohort effects in time series of age-specific marital fertility schedules." Paper presented at the annual meeting of the Population Association of America, Washington, D.C.

Lazarsfeld, P. F. and H. Menzel
1961 "On the relation between individual and collective properties." Pp. 422-440 in A. Etzioni (ed.), Complex Organizations. New York: Holt, Rinehart and Winston.

Mandel, J.
1969 "The partitioning of interaction in analysis of variance." Journal of Research of the National Bureau of Standards, B 73B, No. 4:309-328.

1971 "A new analysis of variance model for non-additive data." Technometrics 13:1-18.

Mare, R. D.
1979 "Social background composition and educational growth." Demography 16:55-71.

Mason, W. M.
1980 "Some statistics and methodology of multi-level analysis." Ann Arbor: Population Studies Center of the University of Michigan.

Mason, W. M. and H. L. Smith
this volume "Age-period-cohort analysis and the study of deaths from pulmonary tuberculosis." [Research Report No. 81-15. Ann Arbor: Population Studies Center of the University of Michigan.]

Nesselroade, J. R. and P. B. Baltes
1974 Adolescent Personality Development and Historical Change: 1970-1972. Monographs of the Society for Research in Child Development, 39 (1, Serial No. 154).

Pullum, T. W.
1977 "Parameterizing age, period and cohort effects: An application to U.S. delinquency rates, 1964-1973." Pp. 116-140 in K. F. Schuessler (ed.), Sociological Methodology 1978. San Francisco: Jossey-Bass.

1980 "Separating age, period and cohort effects in white U.S. fertility, 1920-1970." Social Science Research 9:225-244.

Quandt, R. E.
1974 "A comparison of methods for testing nonnested hypotheses." Review of Economics and Statistics 41:92-99.

Ryder, N. B.
1965 "The cohort as a concept in the study of social change." American Sociological Review 30:843-861.

Smith, D. P.
1981 "A reconsideration of Easterlin cycles." Population Studies 35:247-264.

Tukey, J.
1949 "One degree of freedom for non-additivity." Biometrics 5:232-242.

United States Department of Health, Education and Welfare
1979 Smoking and Health: A Report of the Surgeon General. Washington, D.C.: Government Printing Office.

Winsborough, H. H.
1976 "Discussion." Proceedings of the Social Statistics Section. Pt. 1. American Statistical Association.

Wolfgang, M.
1982 "Delinquency and birth cohort: II." Paper presented at the annual meeting of the American Association for the Advancement of Science, Washington, D.C.

Wolfgang, M. E., R. M. Figlio, and T. Sellin
1972 Delinquency in a Birth Cohort. Chicago: University of Chicago Press.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report #235	2. GOVT ACCESSION NO. A121 925	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Specification and Implementation of Age, Period and Cohort Models		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Stephen E. Fienberg William M. Mason		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Contracts Office Carnegie-Mellon University Pittsburgh, PA 15213		12. REPORT DATE January, 1982 13. NUMBER OF PAGES 74
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

linear models; non-linear interactions; cross-time modelling

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The paper describes models for analyzing cross-time data, using as explanatory variables age, time, and cohort membership, as defined by the period and age at which an individual observation can first enter the period by age data array. The basic "accounting model" for analyzing APC data is described, as is the identification problem created by the linear dependency of age, period, and cohort. By means of a series of examples, various extensions to the APC model are suggested which allow for age-period, age-cohort, and cohort period effects.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)